決定クラスのクラスタリングに基づくルールベース分類モデル(モデリングと最適化の理論)

楠木, 祥文; 乾口, 雅弘

数理解析研究所講究録
1526 巻 2006 年 47-55

決定クラスのクラスタリングに基づくルールベース分類モデル

大阪大学大学院基礎工学研究科
楠木 祥文 (Yoshifumi Kusunoki)
乾口 雅弘 (Masahiro Inuiguchi)
Graduate School of Engineering Science, Osaka Univ.

1. はじめに

近年，データマイニングの分野で，ラフ集合解析 [5] を用いた様々なルール抽出法 [1, 2, 3] が提案されている．これらのルール抽出法では，ルールの結論部はある一つのクラスに帰属することを示すことが多い．しかし，対象 (事例) の分類問題を考えると結論部が必ずしも一つのクラスに対応する必要はなく，ルールの結論部が複数のクラスのいずれかに帰属することを意味しても良い．例えば，三つのクラス D_1，D_2，D_3 に対して，「D_1 or D_2」,「D_2 or D_3」,「D_3 or D_1」を結論部にもつ三つのルールから，未知対象がいずれのクラスに帰属するかを推論することができる．すなわち，未知対象が「D_1 or D_2」で，かつ「D_3 or D_1」であるという推論結果から，その対象が D_1 に帰属することがわかる．一般に，大きなクラスに対するルールは，簡潔になることが期待できるので，D_1，D_2，D_3 それぞれに対応するルールを抽出するよりも「D_1 or D_2」,「D_2 or D_3」,「D_3 or D_1」に対応するルールを抽出した方が効果的でありうる．

この考えに基づいて，本研究ではクラスをグループ化し，そのグループへの帰属を推論するルールを抽出する方法を提案する．この際，どのような方法でクラスをグループ化するかが問題となる．クラスを推論するためには，各クラスがグループの共通集合として表現できなければならない．この問題は複雑な問題であるが，本研究では簡単に，階層的クラスタリングを用いてクラスをグループ化する．階層的クラスタリングにより得られる樹形図を用いたルール抽出法と対象の分類法が提案され，提案法の有用性が数値実験を用いて評価される．

2. ラフ集合とルール抽出

2.1. ラフ集合と決定表

ラフ集合 [5] を用いたデータマイニングで扱われるデータはいくつかの属性を持つ対象の集まりとして与えられる．それを情報表とよび，4 項対 $\langle U, AT, V, \rho\rangle$ で表現される．ここで，U は対象の有限集合，AT は属性の有限集合，V は属性値の集合，$\rho: U \times AT \to V$ は情報関数である．情報表において，属性の部分集合 $A \subseteq AT$ が与えられると，識別不能関係は $R_A \subseteq U \times U$ が次のように定義される．

$$R_A = \{(x,y) \mid \rho(x,a) = \rho(y,a), \forall a \in A\} \tag{1}$$

R_A は，反射性，推移性，対称性を満たすので同値関係である．同値関係 R_A から同値類を次のように定義できる．

$$[x]_{R_A} = \{y \in U \mid (y,x) \in R_A\} \tag{2}$$

ここで，識別不能関係 R_A を用いると，対象の集合 $X \subseteq U$ が与えられたとき，X に確実に含まれている対象の集合を示す下近似 $R_{A*}(X)$ と，X に含まれる可能性がある対象を示す上近似 $R_A^*(X)$ が次のように定められる．

$$R_{A*}(X) \quad = \{x \in U \mid [x]_{R_A} \subseteq X\} = \bigcup\{[x]_{R_A} \mid [x]_{R_A} \subseteq X\} \tag{3}$$

$$R_A^*(X) \quad = \{x \in U \mid [x]_{R_A} \cap X \neq \emptyset\} = \bigcup\{[x]_{R_A} \mid [x]_{R_A} \cap X \neq \emptyset\} \tag{4}$$

$$\tag{5}$$

対 $(R_{A*}(X), R_A^*(X))$ はラフ集合と呼ばれる．ラフ集合理論では，識別不能関係による対象の集合の近似精度や近似の質 [7] が定義され，近似精度や近似の質を損なわない範囲で，できるかぎり条件属性を減らした縮約や条件属性値から決定属性値を導く決定ルールを抽出する方法が議論されている．

情報表の属性集合 AT が条件属性集合 C と決定属性 $\{d\}$ に分割できるとき，情報表を決定表とよぶ．決定表は 4 項対 $T = \langle U, C \cup \{d\}, V, \rho\rangle$ と表される．決定属性による同値類を次のように分割したとき，D_i は決定クラスと呼ばれる．

$$\{[x]_{\{d\}} \mid x \in U\} = \{D_i, i = 1, 2, ..., p\} \tag{6}$$

2.2. MLEM2

本研究では決定ルールを抽出するアルゴリズムとして MLEM2[3] を用いる．MLEM2 は LEM2 の拡張である．LEM2 は決定ルール抽出システム LERS[1] のサブシステムであり，各決定クラスの下近似データを入力すると確実な決定ルール，上近似データを入力すると可能性のある決定ルールを抽出することができる．そのアルゴリズムは，ある基準に従い，抽出するルールの結論部に関連性の強い原始条件 (条件属性と属性値の対) を，条件部を満たすすべての対象が結論部を満たすまで，順次条件部に加えていくアルゴリズムである．いま，ある決定クラスの下近似あるいは上近似を B とし，B に帰属することを導く決定ルールを抽出することを考えよう．$t = (a, v)$ を条件属性 a と属性値 v の対とし，$[t] = \{u \mid u \in U,\ \rho(u, a) = v\}$ と定める．すなわち $[t]$ は条件属性 a の値が v であるという条件を満たす U 内の対象の集合である．T を t の集合とすると，$[T] = \bigcap_{t \in T}[t]$ は T 内すべての t により与えれる条件を満たすような U 内の対象の集合である．$[T] \neq \emptyset$ かつ $[T] \subseteq B$ を満たすとき，B は T に依存するという．B が依存する $T' \supseteq T$ が存在しないとき，T は B の極小集合という．LEM2 では，次の三つの条件を満たす T の集合である $\mathbb{T}$ が求められ，$\mathbb{T}$ に帰属する各 T に対応する条件が求めたい決定ルールの条件部になる．

LEM2 のアルゴリズムは次のように与えられる．

Procedure LEM2
(**input:** a set B, **output:** a single local covering $\mathbb{T}$ of set B)
begin
  $G := B$;
  $\mathbb{T} := \emptyset$;

```
while G ≠ ∅ do begin
  T := ∅;
  T(G) := {t | [t] ∩ G ≠ ∅};
  while T = ∅ or [T] ⊈ B do begin
    † select a pair t ∈ T(G) with the highest priority,  if a tie occurs,  select
      a pair t ∈ T(G) such that |[t] ∩ G| is maximum; if another tie occurs,
      select a pair t ∈ T(G) with smallest cardinality of [t]; if a further tie
      occurs,  select a first pair;
    T := T ∪ {t};
    G := [t] ∩ G;
    T(G) := {t | [t] ∩ G ≠ ∅};
    T(G) := T(G) − T;
  end {while};
  for each t in T do begin
    if [T − {t}] ⊆ B then T := T − {t};
  end {for};
  𝕋 := 𝕋 ∪ {T};
  G := B − ⋃_{T∈𝕋} [T];
  end {while};
for each T in 𝕋 do begin
  if ⋃_{S∈𝕋−{T}} [S] = B then 𝕋 := 𝕋 − {T};
end {for};
end {procedure};
```

LEM2は名義属性を扱うことはできるが，数値属性をもつ決定表に適用すると，条件を満たす対象が少ない決定ルールが得られ，あまり実用的ではない．数値属性を扱うには，LEM2の適用前に離散化する必要がある．入力データに基づき自動的に数値属性の離散化を行う操作を加えることにより，LEM2を修正したMLEM2が提案されている．MLEM2では条件属性が名義属性であるか数値属性であるかを識別し，名義属性である場合には，LEM2と同様に，条件属性と条件属性値の対tを評価する．数値属性である場合には，すべての属性値がソートされ，相隣る二つの値の中点によりカットポイントが生成される．すべてのカットポイントに対して，数値属性がカットポイント以上という条件とカットポイント以下という条件が考えられ，LEM2と同様な方法で条件が評価され，選択される．このよう，MLEM2は数値属性を取り扱うルーチンが加えられたもので，他の操作はLEM2と同様である．

2.3. 未知データの分類

LEM2やMLEM2により決定ルールが抽出されると，これに基づき決定属性値が未知の対象が帰属する決定クラスを推定することが考えられる．未知対象の条件属性データ

が満たす条件をもつ決定ルールのすべてが共通に一つのクラスへの帰属を示唆すれば良いが，異なった決定クラスへの帰属を示す場合がある．一方，未知対象の条件属性データがいずれの決定ルールの条件も満たさない場合もある．これらの場合に対処する方法がLERS[2, 3] の中で与えられており，本研究でもこの方法を用いて対処する．

未知対象の条件属性データが少なくとも一つの決定ルールの条件を満たす場合には，$strength$, $specificity$ の二つに基づく評価基準 $support$ によっていずれの決定クラスに帰属するかを定める．$strength(r)$ はルール r 例となる学習用データの数，$specificity(r)$ はルール r に含まれる (条件属性, 属性値) ペアの数 (条件の長さ) である．このとき，$support(D)$ は未知対象の条件属性データが満足し，かつ，結論部が決定クラス D への帰属を示す決定ルールに関して $strength$ と $specificity$ の積を合計した値であり，次式で定められる．

$$\sum_{matching\ rules\ r\ inferring\ D} strength(r) * specificity(r) \tag{7}$$

この $support$ が最も大きい決定クラスへ対象が分類される．

未知対象の条件属性データがいずれの決定ルールの条件属性を満たさない場合には，一部の条件が満たされるルール (部分合致ルール) が用いられる．部分合致ルール r に対して，ルール r の条件部に含まれる (条件属性, 属性値) ペアの数に対する未知対象が満たす (条件属性, 属性値) ペアの割合である $matching\_factor(r)$ が算出され，次式の値が最も大きい決定クラスに未知対象が分類される．

$$\sum_{partially\ matching\ rules\ r\ inferring\ D} matching\_factor(r) * strength(r) * specificity(r) \tag{8}$$

3. 階層的クラスタリングの適用

本研究では，決定クラスをグループ化するために階層的クラスタリング [8] を用いる．階層的クラスタリングはクラスターと呼ばれる対象のグループをクラスター間の類似度に基づいて逐次結合することにより，階層化された対象のグループを生成する．階層的クラスタリングを適用するためには，決定クラスのクラスター間の類似度を定義する必要がある．

3.1. 類似度

決定クラスのクラスター間の類似度を次の四つのステップを通して定義する．

(1) 対象間の類似度の定義

(2) 対象と決定クラス間の類似度の定義

(3) 決定クラス間の類似度の定義

(4) 決定クラスのクラスター間の類似度の定義

まず，対象間の類似度を定義する．まず，条件属性 $a \in C$ に関する対象間の類似度を定めよう．条件属性には名義属性と数値属性の二つに分類できるので，a が名義属性である

場合と数値属性である場合に分けて，類似度を用いる．名義属性である場合は，$x, y \in U$ とすると，

$$S(x,y;a) = \begin{cases} 1 & \rho(x,a) = \rho(y,a) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

と定める．一方，数値属性である場合は，

$$S(x,y;a) = 1 - \frac{|\rho(x,a) - \rho(y,a)|}{D_a} \tag{10}$$

と定める．ただし，$D_a = \max_{x,y \in U} |\rho(x,a) - \rho(y,a)|$ である．

対象間の類似度はすべての属性に関する類似度平均で定義される．すなわち，

$$S(x,y) = \frac{1}{|C|} \sum_{a \in C} S(x,y;a) \tag{11}$$

と定められる．$S(x,y)$ は x と y がどの程度一致しているかを表している．明らかに，任意の $x, y \in U$ に対し，$S(x,x) = 1$，$S(x,y) = S(y,x)$ である．

次に，対象間の類似度を用いて，対象 x と決定クラス D_i との類似度を次のように定義する．

$$S(x, D_i) = \varphi(\langle S(x,y) \mid y \in B_i \rangle) \tag{12}$$

ただし，B_i は $R_c^*(D_i)$ か $R_{c*}(D_i)$ であり，可能性ルールを抽出したいときは，$R_c^*(D_i)$ を，確実性ルールを抽出したいときは，$R_{c*}(D_i)$ を用いる．$\langle s(x,y) \mid y \in B_i \rangle$ は類似度のマルチ集合で，$\mathcal{M}$ をすべての有限マルチ集合の集合とすると，関数 φ は，$\varphi : \mathcal{M} \to [0,1]$ であり，最大値，平均値，中央値などで定められる．

さらに，$s(x, D_i)$ を用いると，決定クラス間の類似度は次のように定義される．

$$s(D_i, D_j) = \max(\psi(\langle s(x, D_j) \mid x \in B_i \rangle), \psi(\langle s(y, D_i) \mid y \in B_j \rangle)) \tag{13}$$

$\psi : \mathcal{M} \to [0,1]$ であり，φ と同じく最大値，平均値，中央値などの関数である．ここで定めた類似度 $s(D_i, D_j)$ は D_i が D_j に包含される度合を表している．この類似度は，各対象の条件属性パターンが似た対象を含むクラスをまとめれば，決定ルールが簡潔になるだろうというヒューリスティックスに基づいている．

本研究では，確実性ルールを抽出したいので，式 (12)，(13) の B_i，B_j として，決定クラスの下近似 $C_*(D_i)$，$C_*(D_j)$ を採用する．

最後に決定クラスのクラスター間の類似度 $S(G_i, G_j)$ は次のように定義される．

$$S(G_i, G_j) = S\left(\bigcup_{D_k \in G_i} D_k, \bigcup_{D_l \in G_j} D_l \right) \tag{14}$$

与えられた決定表に矛盾が存在しない場合，φ と ψ がともに最大値で定められるとき，階層的クラスタリングの最短距離法と一致し，φ と ψ がともに最小値で定められる場合，最長距離法と一致する．

4. 樹形図を用いたルール抽出と分類

決定クラスのクラスター間の類似度が定義できたので，階層的クラスタリング法を適用して決定クラスの類似度を示す樹形図を求めることができる．樹形図に従ったルール抽出法について述べる．未知対象の識別を考えると，樹形図の各分枝点において，対象がいずれのグループに帰属するかを決定する必要がある．その決定を行うルールを抽出する方法として，次の二つが考えられる．一つは，分枝点の両側の枝に対して，ルールを抽出する方法であり，もう一つは，分枝点において片側の枝に対してのみルール抽出を行う方法である．後者の方法は，対象数が小さいグループについてルール抽出を行う方法と大きいグループついてルールを抽出を行う方法の二通りに分けられる．結果として，樹形図を用いた決定ルール抽出法として，次の三つのモードが考えられる．

Accuracy Mode 樹形図の各分枝点において，両側の枝に対して決定ルールの抽出を行う方法である．他の方法より決定ルール抽出アルゴリズムを多く適用するので，計算量が多く，決定ルールの数も多くなる．その反面，より正確な識別を期待できる．

Simplicity Mode 樹形図の各分枝点において，含まれている訓練用データ数が大きいグループの枝に対してのみ，決定ルールの抽出を行う方法であり．より多くのデータに対する決定ルールの方が簡潔になるであろうという考えに基づいている．対象が抽出した決定ルール群のいずれかの条件を満たせば，ルール抽出を行ったグループに帰属すると判定され，そうでなければ，他方のグループに帰属すると判定される．

Majority Mode 樹形図の各分枝点において，含まれている訓練用データ数が小さいグループの枝に対してのみ決定ルールの抽出を行う方法である．ルールの条件に合致した対象はそのグループに帰属すると判定され，合致しないものはすべて他方のグループに帰属すると判定される．大きいグループをデフォルトであるとみなす考えに基づいた方法である．

本研究では，グループに含まれる決定クラスの和集合の下近似に対して MLEM2 を適用し，決定ルールを抽出する．また， Accuracy Mode では完全な分類を行うため，先に述べた未知データの分類法を用いる．

5. 数値実験

提案法の有用性を評価するために数値実験を行った．提案法では，関数 φ,ψ と三つのモードがあり，様々な組合せが考えられる．ここでは，三つの組合せについて実験を行った．三つの組合せはいずれも $\varphi=\psi$ とし，Accuracy Mode を採用している．φ，ψ は第一の方法では，平均を、第二の方法では，最大値を，第三の方法では，最小値を用いている．比較のため，MLEM2 を用いて単純に決定ルールを抽出する従来法と決定木を生成する C4.5[6] についても実験を行う．また，階層的クラスタリングで用いた類似度の有効性を評価するため，類似度を用いずにランダムにクラスターを結合して樹形図を得る方法についても数値実験を行った．この方法でも樹形図の各分枝点で Accuracy Mode を適用した．

UCI Machine Learning Repository[4] より入手した実験用データを 5-fold cross validation を用い訓練用データを検証用データに分割し，訓練用データから決定ルールを生成

表 1: 実験用データ

データ名	対象数	条件属性数	決定クラス数
dermatology	360	34	6
ecoli	336	7	8
glass	214	10	7
hayes-roth	160	4	3
segmentation	210	19	7
iris	150	4	3
wine	178	13	3
zoo	101	16	7

し，それらを用いて検証用データの分類結果を評価する実験を5回行った．評価基準として，抽出した決定ルールの数 (NUM)，決定ルールに含まれる条件数の平均値つまり決定ルールの平均長 (LEN)，決定ルール群または決定木に含まれる条件数の総和 (SIZE)，分類の誤識別率 (INC)，実行時間 (TIME) の五つを用いる．10 回の 5-fold cross validation の適用により，各評価基準に対し 50 個の評価値を得るが，表 2 の実験結果はその平均値を示している．本研究で用いた七つの実験用データ集合の対象数，条件属性数，決定クラス数は表 1 のようになっている．

提案法と通常の MLEM2 および C4.5 と比較すると，φ，ψ のいずれの決定においても，提案法のルール数 (NUM) は MLEM2 に比らべ増加しているが，ルールの長さ (LEN) は短くなっている．条件数 (SIZE) に注目すると，ほとんどのデータ集合に対して，提案法と MLEM2 は同じくらいであるが，dermatology と hayes-roth については提案法の方が小さくなっており，やや簡潔になっている．一方，C4.5 と比較すると C4.5 の方が SIZE が小さく，簡潔になっている．分類の誤識別率 (INC) について見ると，提案法の方が，ほとんどのデータ集合で，MLEM2 より良くなっているが，C4.5 と比較すると，dermatology, hayes-roth，iris，wine，zoo の五つのデータのみで良い結果が得られている．

次に，ランダムに樹形図を生成した方法 (Random) と提案法を比較すると，NUM，LEN，SIZE とも glass と hayes-roth を除き，提案法の方が小さく，得られたルール群が簡潔になっていることが分かる．これより，類似度を用いたクラスタリングが有効であることが分かる．また，誤識別率 (INC) も概ね提案法の方が小さく良い結果を示している．

φ，ψ の決定が異なる三つの提案法間で比較を行うと，(min,min) が最も悪い結果を示す場合には，他の二つよりかなり悪いものとなる．(max,max) は (min,min) ほど悪くはないが，MLEM2 や Random より識別精度が下がることがある．以上より，(mean,mean) は極端に悪くなることがなく安定しており，三つの提案法の中では最も良い方法であると考えられる．

最後に実行時間 (TIME) をみると，実験に用いたデータ集合の対象数が少ないため，提案法と MLEM2 に大きな差が出ていない．

6. おわりに

階層的クラスタリングの適用により決定クラスを階層的にグループ化し，樹形図に従ったルール抽出法を提案した．数値実験を通して，決定クラスをグループ化することによ

表 2: 実験結果

(1) dermatology の結果

手法	NUM	LEN	SIZE	INC	TIME
(mean,mean)	23.2	2.3	52.2	5.5%	0.02sec.
(max,max)	22.9	2.2	51.0	4.9%	0.02sec.
(min,min)	25.6	2.3	59.2	5.3%	0.02sec.
MLEM2	18.3	3.8	68.7	11.8%	0.02sec.
Random	28.7	2.6	74.3	8.9%	0.03sec.
C4.5	–	–	30.1	6.6%	0.01sec.

(2) ecoli の結果

手法	NUM	LEN	SIZE	INC	TIME
(mean,mean)	43.6	2.8	121.4	19.1%	0.04sec.
(max,max)	46.3	2.7	125.2	19.3%	0.05sec.
(min,min)	46.1	2.9	132.8	20.0%	0.05sec.
MLEM2	31.3	3.9	122.5	22.5%	0.05sec.
Random	47.9	2.9	139.4	21.0%	0.06sec.
C4.5	–	–	31.8	18.3%	0.01sec.

(3) glass の結果

手法	NUM	LEN	SIZE	INC	TIME
(mean,mean)	34.6	3.1	106.7	33.2%	0.05sec.
(max,max)	34.7	3.0	104.2	32.0%	0.06sec.
(min,min)	36.2	3.3	119.9	36.2%	0.08sec.
MLEM2	24.0	4.3	102.2	34.3%	0.05sec.
Random	36.5	3.2	117.7	34.1%	0.07sec.
C4.5	–	–	40.9	33.0%	0.01sec.

(4) hayes-roth の結果

手法	NUM	LEN	SIZE	INC	TIME
(mean,mean)	15.6	2.4	37.7	15.1%	0.00sec.
(max,max)	18.5	2.6	48.6	17.8%	0.01sec.
(min,min)	15.6	2.4	37.7	15.1%	0.00sec.
MLEM2	14.6	3.5	50.3	20.6%	0.01sec.
Random	17.7	2.5	45.2	17.1%	0.00sec.
C4.5	–	–	22.0	17.5%	0.00sec.

(5) segmentation の結果

手法	NUM	LEN	SIZE	INC	TIME
(mean,mean)	19.6	2.0	39.9	13.5%	0.04sec.
(max,max)	19.5	2.0	39.0	12.7%	0.03sec.
(min,min)	20.5	2.1	43.8	14.0%	0.06sec.
MLEM2	12.0	3.2	38.4	15.1%	0.03sec.
Random	23.1	2.5	56.6	16.4%	0.09sec.
C4.5	–	–	23.7	12.7%	0.02sec.

(6) iris の結果

手法	NUM	LEN	SIZE	INC	TIME
(mean,mean)	8.1	2.0	16.6	4.4%	0.01sec.
(max,max)	8.1	2.0	16.6	4.4%	0.01sec.
(min,min)	8.1	2.0	16.2	5.7%	0.00sec.
MLEM2	7.2	2.4	17.3	5.5%	0.00sec.
Random	8.6	2.0	17.6	4.7%	0.00sec.
C4.5	–	–	7.5	5.3%	0.00sec.

(7) wine の結果

手法	NUM	LEN	SIZE	INC	TIME
(mean,mean)	5.8	2.3	13.4	5.6%	0.01sec.
(max,max)	5.9	2.3	13.6	5.8%	0.01sec.
(min,min)	5.8	2.3	13.5	5.7%	0.01sec.
MLEM2	4.5	3.2	14.1	7.5%	0.01sec.
Random	6.3	2.5	15.9	5.6%	0.01sec.
C4.5	–	–	9.5	7.7%	0.01sec.

(8) zoo の結果

手法	NUM	LEN	SIZE	INC	TIME
(mean,mean)	14.7	1.2	17.4	5.5%	0.01sec.
(max,max)	13.9	1.1	15.7	5.3%	0.00sec.
(min,min)	14.8	1.2	17.8	6.3%	0.01sec.
MLEM2	9.1	2.1	19.2	6.3%	0.00sec.
Random	18.2	1.4	25.4	7.4%	0.00sec.
C4.5	–	–	16.6	8.4%	0.00sec.

り分類精度に改善が見られることを示すとともに，階層的クラスタリングに用いた類似度の有効性を確認した．今後の課題として，階層的クラスタリングとは別の方法で決定クラスのグループ化し，そのグループを用いたルール抽出法の提案などがあげられる．

参考文献

[1] Grzymala-Busse, J. W.: LERS – A system for learning from examples based on rough sets, *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory* edited by R. Słowínski, Kluwer Academic Publishers, Dordrecht, pp.3-18, 1992.

[2] Grzymala-Busse, J. W. and Stefanowski, J.: Three Discretization Methods for Rule Induction, *International Journal of Intelligent Systems*, **16**, 2001, 29-38.

[3] Grzymala-Busse, J. W.: MLEM2 - Discretization During Rule Induction, *International Conference on Intelligent*, 2003, 499-508.

[4] Newman, D. J., Hettich, S., Blake, C. L., Merz, C. J.: UCI repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html], 1998.

[5] Pawlak, Z.: Rough sets, *Internat. J. of Inform. & Comput. Sci.*, Vol.11, No.5, pp.341-356, 1982.

[6] Quinlan, J. R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[7] 乾口 雅弘:ラフ集合による情報の解析, システム/制御/情報, Vol.49, No.5, pp.165-172, 2005.

[8] 宮本 定明：クラスター分析入門, pp.88-105, 森北出版, 1999.